Wavelet Speech Enhancement based on the Teager Energy Operator

Mohammed Bahoura and Jean Rouat*

ERMETIS, DSA, Université du Québec à Chicoutimi,
Chicoutimi, Québec, G7H 2B1, Canada.

Abstract

We propose a new speech enhancement method based on the time adaption of wavelet thresholds. The time dependence is introduced by approximating the Teager Energy of the wavelets coefficients.

This technique does not require an explicit estimation of the noise level or of the *a priori* knowledge of the SNR, which is usually needed in most of the popular enhancement methods. Performance of the proposed method are evaluated on speech recorded in real conditions and with artificial noise.

SPL.SA.1.5

*Corresponding author. email: jrouat@uqac.uquebec.ca

1 Introduction

During the past decade, wavelet transforms have been applied to various research areas. Their applications include signal and image denoising, compression, detection, and pattern recognition.

Wavelet shrinkage is a simple denoising technique based on the thresholding of the wavelet coefficients. The estimated threshold is supposed to define the limit between the wavelet coefficients of the noise and those of the target signal. Unfortunately it is not always possible to separate the components corresponding to the target signal from those of noise by a simple thresholding. For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients not only suppresses additional noise but also some speech components like unvoiced ones. Consequently, the perceptive quality of the filtered speech will be greatly affected.

However, the wavelet transform combined with other signal processing tools has been proposed for speech enhancement. They include the Wiener filtering in the wavelet domain [1], wavelet filter bank for spectral subtraction [2] or coherence function [3, 4].

We propose a novel approach for wavelet speech enhancement. Unlike conventional denoising wavelet methods, the discriminative threshold in various scales is time adapted in function of speech components. The proposed technique is tested on noisy speech recorded in real environments. Obtained results are closely similar to those from Ephraim and Malah Filter (EMF) [5, 6]. Furthermore, the proposed method does not require *a priori* any knowledge of the SNR.

2 Theory

2.1 Wavelet denoising

Wavelet transform has recently emerged as a powerful tool for removing noise from signal and image. Donoho and Johnston proposed their original method, which proceeds by thresholding wavelet coefficients [7, 8]. They attempt to recover a signal s(t) from noisy data x(t)

$$x_i = s_i + b_i \qquad i = 1, \ldots, N \tag{1}$$

where b_i represents a gaussian white noise.

This algorithm can be summarized in three steps

- Wavelet transform of the noisy signal,
- Thresholding the resulting wavelet coefficients,
- Inverse wavelet transform to obtain the denoised signal.

Donoho and Johnstone [8] proposed a universal threshold λ for removing added white noise

$$\lambda = \sigma\sqrt{2\log(N)} \tag{2}$$

$$\text{with } \sigma = MAD/0.6745 \tag{3}$$

where σ is the noise level. MAD is the Median Absolute Deviation, estimated in the first scale. In the wavelet packets case, the threshold becomes

$$\lambda = \sigma\sqrt{2\log(N\log_2 N)} \tag{4}$$

The soft thresholding function is defined as

$$T_S(\lambda, w_k) = \begin{cases} sgn(w_k)(|w_k| - \lambda) & \text{if } |w_k| \geq \lambda \\ 0 & \text{if } |w_k| < \lambda \end{cases} \tag{5}$$

where w_k represents the wavelet coefficients.

2.2 Teager energy operator

The Teager Energy Operator (TEO) is a powerful nonlinear operator proposed by Kaiser [9], capable to extract the signal energy based on mechanical and physical considerations. It has been successfully used in various speech applications [10, 11, 12, 13, 14]. For a bandlimited digital signal $x(n)$, this operator can be approximated by

$$\Psi_d[x(n)] = [x(n)]^2 - x(n+1)x(n-1) \tag{6}$$

3 New enhancement method

The proposed speech enhancement method is based on the time adaptation of the wavelet threshold. Fig. 1 explains schematically this algorithm for a short noisy sentence (Fig. 1(a)).

3.1 Wavelet packet analysis

For a given level j, the wavelet packets transform WP decomposes the noisy signal $x(n)$ into 2^j subbands corresponding to wavelet coefficient sets $w^j_{k,m}$. For this application, we fix $j = 4$.

$$w^j_{k,m} = WP\{x(n), j\} \qquad n = 1, \ldots, N \tag{7}$$

In other words, $w^4_{k,m}$ defines the m^{th} coefficient of the k^{th} subband. Where $m = 1, ..., N/2^4$

and $k = 1, ..., 2^4$. For example, Fig. 1(b) represents the wavelet coefficient set $w^4_{5,m}$.

3.2 Teager energy approximation

The discrete–time TEO is applied to the resulting wavelet coefficients $w^4_{k,m}$

$$t^4_{k,m} = \Psi_d[w^4_{k,m}] \qquad k = 1, \ldots, 16 \tag{8}$$

This operation enhances the discriminability of speech coefficients among those of noise (Fig. 1(c)).

3.3 Masks Construction

For each subband coefficients, an initial mask is obtained by smoothing the TEO coefficients (Fig. 1(d))

$$\mathrm{M}^4_{k,m} = t^4_{k,l} * h_k(m) \tag{9}$$

where h_k is an IIR lowpass filter (2^{th} order).

3.4 Threshold modulation criterion

Ideally, the standard threshold should be adapted only for speech frames and kept unchanged for non speech ones. The speech presence is interpreted by a significant contrast between peaks and valleys of M^4_k, while its absence is observed with a weaker contrast (smoother masks). To distinguish these frames, we define a parameter S^4_k named *offset*, that estimates the valleys level. It is given by the abscissa of the maximum of the amplitude distribution H of the corresponding mask $M^4_{k,m}$, and is estimated over the analyzed frame.

$$S^4_k = abscissa[max(H(M^4_{k,m}))] \tag{10}$$

This parameter is close to 0 for speech frames and close to 1 for noise ones. If S^4_k is below the discriminatory value of $0.35max(M_{k,m}^4)$ then, we modulate the threshold, else it remains unchanged.

3.5 Mask processing for the time adapted threshold

The modulated threshold must be adapted to the speech waveform independently of its energy evolution. In this case, the difference between local maxima must be reduced. We proceed by suppressing the *offset* and normalization, before applying a root power function

$$M_{k,m}'^4 = [\frac{M_{k,m}^4 - S_k^4}{max(M_{k,m}^4 - S_k^4)}]^{\frac{1}{8}} \tag{11}$$

$M_{k,m}'^4$ is reported in Fig. 1(e) for $k = 5$.

3.6 Time adapted threshold

For a given subband k, we define the time adapted threshold as

$$\lambda_{k,m} = \lambda(1 - \alpha M_{k,m}'^4) \tag{12}$$

where λ is the standard threshold (Equation 4) and α an ajustment parameter ($\alpha = 1$).

Fig. 1(f) represents the standard threshold λ (dashed line) and the resulting time adapted threshold $\lambda_{k,m}$ (continuous line) for the subband $k = 5$.

3.7 Thresholding

The soft thresholding (Equation 5) is then applied to the wavelet packet coefficients (Fig. 1(g))

$$\widehat{w}_{k,m}^4 = T_S(\lambda_a, w_{k,m}^4) \tag{13}$$

where λ_a is the threshold corresponding to the analyzed frame.

$$\lambda_a = \begin{cases} \lambda_{k,m} & \text{if } S_k^4 \leq 0.35max(M_{k,m}^4) \\ \lambda & \text{if } S_k^4 > 0.35max(M_{k,m}^4) \end{cases} \tag{14}$$

3.8 Inverse transformation

The enhanced signal (Fig. 1(h)) is synthetized with the inverse transformation WP^{-1} of the resulting wavelet coefficients

$$\widehat{s}_n = WP^{-1}\{\widehat{w}_{k,m}^4, j\} \tag{15}$$

4 Results and discussion

The proposed method is evaluated using natural speech corrupted by white noise and speech recorded in real environments (in a sawmill, in aircraft cockpit, and street). The speech data are sampled at 8 kHz.

A clean sentence from the TIMIT database is corrupted by white noise for various SNR ranging from -10 to 20 dB. A sliding overlapping window of 32 ms is used in the Ephraim and Malah Filter. While in our method the whole speech frame is filtered once by the WPF (Wavelet Packets Filter). However, a minimum length of the frame is required for the WPF. It must be sufficiently large to include simultaneously speech and silence.

Results of the enhancement obtained by the EMF and the proposed method WPF are reported in Table 5.

Table 5 shows that the proposed method is well suited for very strong noise with a SNR ranging from -10 to 10 dB and with better performance than the EMF.

Examples of enhancement for white noise corrupted speech and aircraft recorded speech (real environment) are respectively reported in Fig. 2 and 3. We recall that the EMF algorithm requires an explicit estimation of the noise level or the *a priori* knowledge of the SNR, which is not necessary for our system.

5 Conclusion

To our knowledge, the proposed method is one of the first successful applications of the wavelet thresholding method for speech enhancement. The discriminatory threshold is time–adapted to the speech waveform, unlike applying the time–constant standard threshold. For very strong noise, the proposed method yields higher SNR than the Ephraim and Malah Filter.

References

[1] D. Mahmoudi, “A microphone array for speech enhancement using multiresolution wavelet transform,” in *Proc. Of Eurospeech’97*, Rhodes, Greece, September 1997, pp. 339–342.

[2] T. Gulzow, A. Engelsberg, and U. Heute, “Comparison of a discrete wavelet transformation and nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement,” *Signal Processing*, vol. 64, pp. 5–19, 1998.

[3] J. Sika and V. Davidek, “Multi-channel noise reduction using wavelet filter bank,” in *EuroSpeech’97*, 1997.

[4] D. Mahmoudi and A. Drygajlo, “Combined wiener and coherence filtering in wavelet domain for microphone array speech enhancement,” in *ICASSP*, Seattle, USA, 1998, pp. 385 –388.

[5] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[6] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[7] D.L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[8] D.L. Donoho and I.M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[9] J.F. Kaiser, “On a simple algorithm to calculate the ’energy’ of a signal,” in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, 1990, pp. 381–384.

[10] P. Maragos, T. Quatieri, and J.F. Kaiser, “Speech non-linearities, modulation and energy operators,” in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Toronto, 1991, pp. 421–424.

[11] J.F. Kaiser, “Some useful properties of teager’s energy operators,” in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, April 1993, vol. 3, pp. 149–152.

[12] J. Rouat, “Nonlinear operators for speech analysis,” in *Visual representations of speech signals*, M. Cooke, S. Beet, and M. Crawford, Eds. 1993, pp. 335–340, J. Wiley and Sons.

[13] J. Rouat, Y.C. Liu, and D. Morissette, “A pitch determination and voiced/unvoiced decision algorithm for noisy speech,” *Speech Communication*, vol. 21, pp. 191–207, 1997.

[14] F. Jabloun, A.E. Cetin, and E. Erzin, “Teager energy based feature parameters for speech recognition in car noise,” *IEEE Signal Processing Letters*, vol. 6, no. 10, pp. 259–261, October 1999.

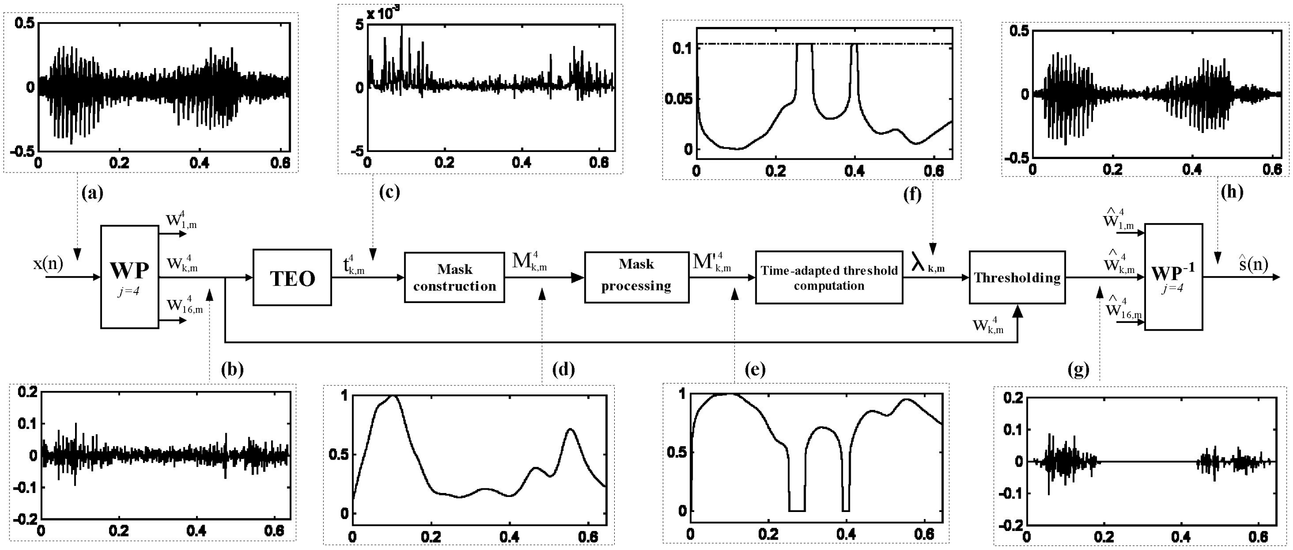
Table I. SNR tests for white noise corrupted speech

Fig. 1. Speech enhancement diagram using time–adapted thresholding in the wavelet packet domain

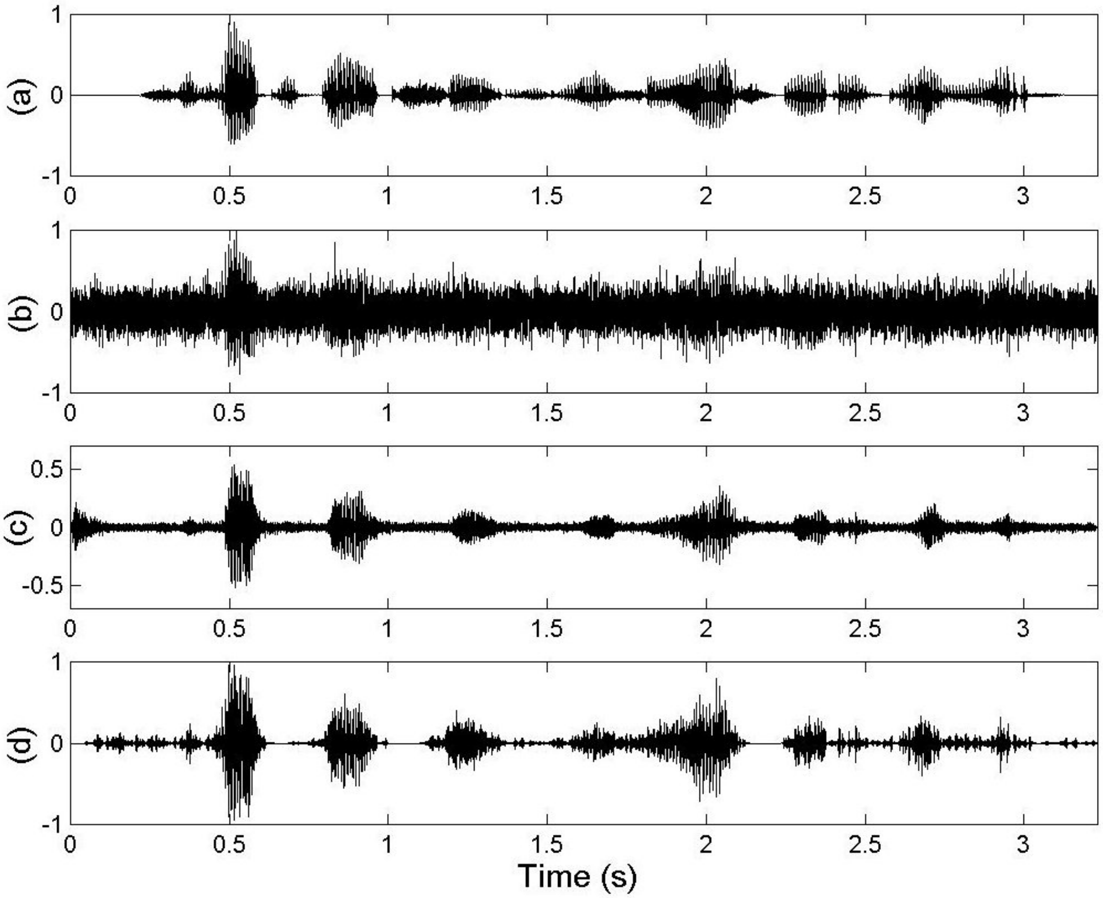
Fig. 2. Speech enhancement results: a) clean signal, b) noisy version (SNR=-5dB), c) enhanced with EMF, and d) enhanced with WPF

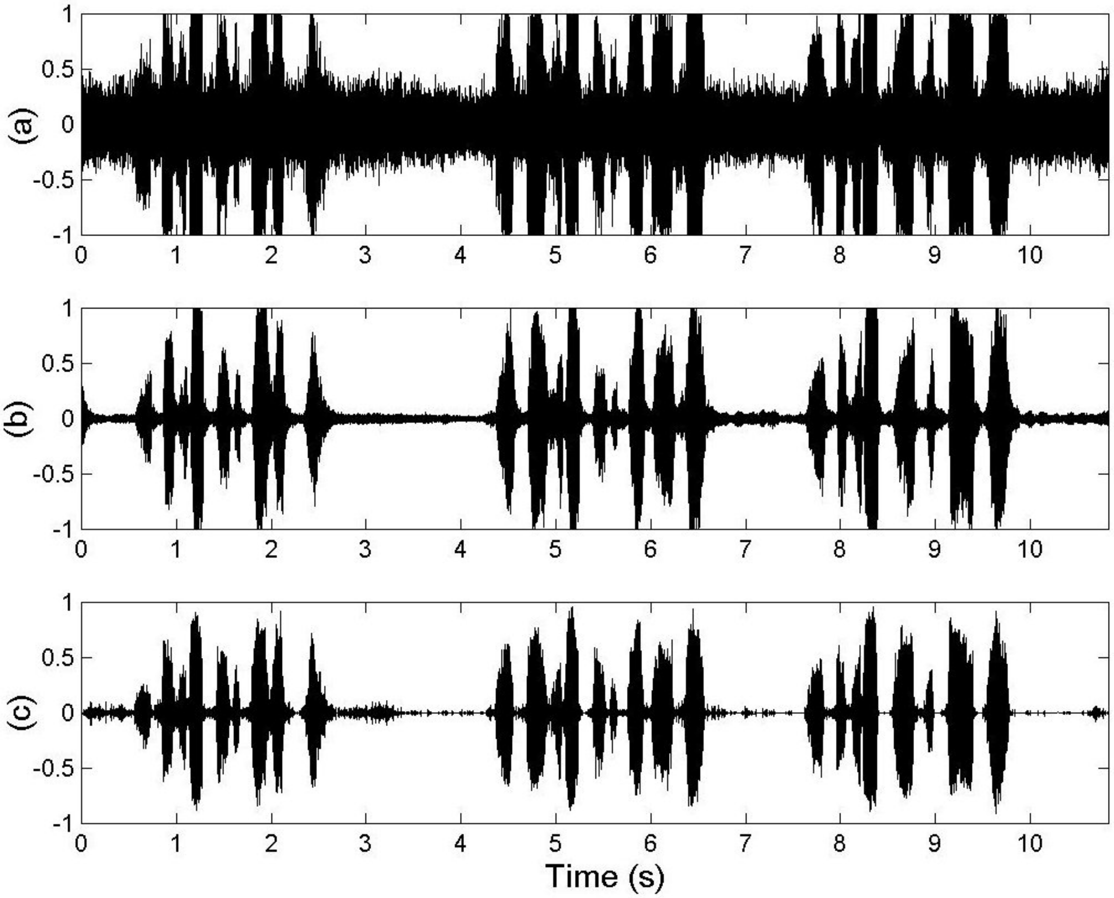
Fig. 3. Speech enhancement results: a) noisy speech recorded in an aircraft cockpit, b) enhanced speech with the EMF, and c) enhanced with the WPF

Unprocessed (dB)	EMF (dB)	WPF (dB)
-10	1.68	2.41
-5	4.05	4.04
0	6.40	6.51
5	8.98	9.23
10	11.91	12.10
15	15.54	14.47
20	19.67	16.47



(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
x(n)
WP
j=4
$W_{1,m}^4$
$W_{k,m}^4$
$W_{16,m}^4$
TEO
$t_{k,m}^4$
Mask construction
$M_{k,m}^4$
Mask processing
$M'^4_{k,m}$
Time-adapted threshold computation
$\lambda_{k,m}$
Thresholding
$W_{k,m}^4$
$\hat{W}_{1,m}^4$
$\hat{W}_{k,m}^4$
$\hat{W}_{16,m}^4$
WP^{-1}
j=4
$\hat{s}(n)$
0.5
0
-0.5
0.2
0.1
-0.1
-0.2
x 10^{-3}
5
-5
1
0.05
0.4
0.6


(a)
1
0
-1
(b)
1
0
-1
(c)
0.5
0
-0.5
(d)
1
0
-1
0
0.5
1
1.5
2
2.5
3
Time (s)



(a)
(b)
(c)
1
0.5
0
-0.5
-1
0
1
2
3
4
5
6
7
8
9
10
Time (s)